La condensation de l'information

Michel Bellot-Antony, Gabriel G. Bès, Dany Hadjadj, Régine Pouzet, Nicole Rousseau-Payen

La condensation de l’information

Michel Bellot-Antony, Gabriel G. Bès, Dany Hadjadj, Régine Pouzet et Nicole Rousseau-Payen
Groupe de recherches sur la condensation de l’information en langue naturelle (CILN)

Condenser, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 1-9.

Résumé

Ce texte présente les programmes de recherche du CILN, sur la contraction de texte, les langages documentaires et la codification économique du langage.

Voir aussi

Michel Bellot-Antony, Gabriel G. Bès, Dany Hadjadj, Régine Pouzet et Nicole Rousseau-Payen. « La problématique de la contraction de texte. » *Condenser*, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 13-44.

Michel Bellot-Antony, Gabriel G. Bès, Dany Hadjadj, Régine Pouzet et Nicole Rousseau-Payen. « La contraction de texte. » *Condenser*, Adosa, Clermont-Ferrand, janvier 1981, n° 2, p. 5-38.

Michel Bellot-Antony, Gabriel G. Bès, Dany Hadjadj, Régine Pouzet et Nicole Rousseau-Payen. « La contraction de texte. Les différents types d'information. » *Condenser*, Adosa, Clermont-Ferrand, avril 1982, n° 3, p. 33-81.

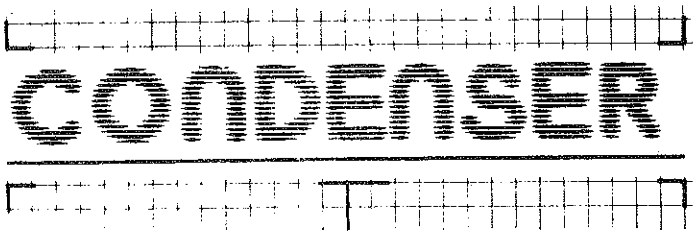
Ryszard Zuber. **«** Relations sémantiques et résumé. » *Condenser*, Adosa, Clermont-Ferrand, avril 1982, n° 3, p. 83-94.

Michel Bellot-Antony et Gabriel G. Bès. « Les différents types d'information et la contraction de texte. » *Condenser*, Adosa, Clermont-Ferrand, février 1983, n° 4, p. 5-46.

Dany Hadjadj. « La contraction de texte au baccalauréat : qu’en pensent les enseignants ? » *Condenser*, Adosa, Clermont-Ferrand, février 1983, n° 4, p. 47-65.

Michel Bellot-Antony. « Bibliographie commentée (suite). » *Condenser*, Adosa, Clermont-Ferrand, février 1983, n° 4, p. 67-74.

Dominique Le Roux. « Les conditions sémantiques de l’activité résumante dans les textes de sciences sociales et humaines. » *Condenser*, Adosa, Clermont-Ferrand, février 1983, n° 4, p. 75-93.



CAHIERS DU GROUPE DE RECHERCHES

SUR LA

CONDENSATION DE L'INFORMATION EN LANGUE NATURELLE

N° 1

FÉVRIER 1980

ADOSA

CLERMONT-FERRAND

2ème tirage: juin 1980.

A V A N T - P R O P O S

Ce bulletin est né de la constatation qu'entre la recherche en train de se faire et les circuits traditionnels de publication le hiatus est trop grand. Diverses enquêtes ont montré qu'un délai, qui dépasse souvent deux ans dans certaines disciplines, sépare le moment où un travail de recherche atteint le stade final de la rédaction et celui où il est à la disposition des lecteurs dans une revue ou sous la forme d'un volume. Il faut quelques mois de plus pour qu'il soit signalé dans les bibliographies courantes ou enregistré dans les bases de données et qu'il puisse être ainsi repéré facilement par l'ensemble de la communauté scientifique.

Il est alors trop tard, bien souvent, pour que les réactions suscitées par ce travail influencent positivement, en retour, le chercheur ou l'équipe de recherche, qui sont déjà engagés dans d'autres directions. Le dialogue n'a pu s'établir auparavant qu'à travers des circuits confidentiels et au hasard des rencontres personnelles et des réunions scientifiques. A l'appui de ces échanges circule toute une littérature souterraine, difficile à identifier, à collecter et à citer. Cette situation n'est satisfaisante ni pour les auteurs dont les droits sont mal protégés et les travaux tardivement reconnus, ni pour les autres chercheurs qui sont imparfaitement informés, ou qui, quand ils le sont, ne peuvent utiliser et citer commodément des résultats dont ils ont eu connaissance par des voies indirectes; elle ne l'est pas davantage pour l'avancement des idées scientifiques puisque c'est souvent dans cette littérature souterraine, inaccessible au plus grand nombre, qu'apparaissent des démonstrations qu'on a tendance ensuite à considérer comme acquises, sans qu'ait été mis à la portée de tous les chercheurs le moyen de les vérifier.

Pour remédier à cet état de choses dans le domaine qui est le nôtre, et selon une formule déjà tentée ailleurs avec succès, nous proposons

C O N D E N S E R .

Condenser n'est pas une revue; c'est l'organe du Groupe de Recherches sur la Condensation de l'Information en Langue Naturelle (C.I.L.N.), qui veut en faire un instrument de communication scientifique plus souple qu'une revue, le support d'un dialogue permanent entre équipes et chercheurs intéressés par les problèmes que posent la condensation de l'information véhiculée par les textes en langue naturelle et, plus généralement, la représentation de l'information.

Condenser recueille soit des articles entièrement élaborés, soit des versions provisoires d'articles, ou de chapitres à intégrer à des ouvrages en gestation, ainsi que des fiches de travail, des indications bibliographiques, des comptes rendus, des informations diverses. Ces contributions ne se plieront pas nécessairement aux formes imposées par la rhétorique habituelle de la communication scientifique. Les hésitations doivent pouvoir s'y exprimer ainsi que les hypothèses de travail en cours de vérification. Les opinions et les résultats pourront donc être modifiés et améliorés en fonction des progrès de la recherche et de la confrontation avec les lecteurs.

Ces Cahiers souhaitent accueillir, à côté des travaux des chercheurs du C.I.L.N., des articles proposés par des auteurs extérieurs, qui, conformément à la formule que nous avons choisie, resteront libres de reprendre ultérieurement, pour une autre publication, les textes parus dans Condenser. Tous les ouvrages reçus qui intéressent le champ de recherche du C.I.L.N. seront signalés et les plus importants feront l'objet d'un compte rendu.

Afin d'adapter au mieux le rythme de production aux besoins et aux possibilités des chercheurs, il a été décidé de ne pas fixer de manière rigide le nombre de fascicules à faire paraître chaque année (il y aura sans doute, en moyenne, deux numéros par an). En conséquence, les abonnements seront pris pour un nombre donné de numéros et non pour une durée déterminée. La présentation matérielle de ces Cahiers pourra et devra être améliorée pour les numéros suivants, mais en respectant, dans le choix des méthodes de fabrication, les exigences prioritaires qui sont celles de la rapidité et de l'efficacité.

LA CONDENSATION DE L ' INFORMATION

Intuitivement, tout usager d'une langue a le sentiment qu'un certain nombre de textes en langue naturelle sont redondants et très longs, qu'ils manquent de densité, et que, par conséquent, ils sont susceptibles d'être raccourcis, résumés, contractés, condensés. Condenser un texte, c'est s'attacher à exprimer son contenu sous une forme plus économique, adaptée à d'autres conditions de communication que celles du texte initial. A partir de ces intuitions, très largement partagées, le C.I.L.N. s'est fixé pour objectif l'étude d'une problématique de la condensation de l'information fournie au moyen d'une langue naturelle.

Tout texte en langue naturelle véhicule de l'information. Celle-ci est conditionnée par une pluralité de facteurs : le système linguistique sous-jacent au texte (qu'on le nomme "grammaire", "code" ou "langue"), la structuration du texte et des énoncés qui le composent, les mécanismes qui produisent et utilisent le texte, la connaissance des conditions de l'énonciation, les lois rhétoriques, mais aussi la pré-connaissance du "monde", en particulier la pré-connaissance "référentielle" du monde dont parle le texte et ce que l'on sait, ou croit savoir, des connaissances de celui auquel le texte est destiné. Cette pluralité de facteurs qui conditionnent l'information véhiculée par le texte va fortement relativiser des notions telles que "information essentielle", "sous-entendu", "dire la même chose de manière différente", notions qui semblent pourtant intuitivement claires. Ainsi, on peut admettre que les énoncés (1) et (2) ci-dessous disent la même chose pour tous les francophones; mais, parmi eux, seule une catégorie particulière estimera que (3) et (4) véhiculent une même information :

(1) Jacques n'a pas cassé la vitre bien qu'il ait claqué la porte-fenêtre.

(2) Jacques a claqué la porte-fenêtre, mais il n'a pas cassé la vitre.

(3) Ce sujet présente une rétraction des ischios jambiers.

(4) Ce sujet présente un allongement du quadriceps.

Condenser, n°1, 1980, 5-9.

Considérons par ailleurs le texte suivant :

(5) Un calme apparent règne au volcan de la Soufrière de l'île de Saint-Vincent. L'évacuation des villages situés à proximité dans le nord de l'île se poursuit cependant. Les experts restent en effet inquiets. Ils ne savent pas si ce calme marque la fin de l'alerte ou annonce une éruption prochaine. Les quelque quinze mille personnes qui ont déjà été évacuées dans le sud de l'île ne regagneront pas leurs villages avant au moins un mois.

Le Populaire du Centre du 17.4.79

Il est difficile - sinon impossible - sans tenir compte de facteurs qui ne sont pas dans le texte, mais qui sont autour du texte, de décider lequel de (5.1) ou de (5.2) présentés ci-après est la meilleure condensation aux 2/5èmes du texte (5), (5.1) étant orienté vers la situation géologique et (5.2) vers la situation humaine.

(5.1) Le calme qui règne au volcan de la Soufrière (île de Saint-Vincent) inquiète les experts qui se demandent s'il n'annonce pas une éruption prochaine; l'évacuation des habitants se poursuit donc.

(5.2) L'évacuation du Nord de l'île de Saint-Vincent se poursuit, le calme du volcan inquiétant les experts; le retour des habitants évacués au Sud ne se fera pas avant un mois.

Pourquoi a-t-on besoin d'un texte condensé? Deux types de réponse semblent possibles. D'une part, on veut obtenir un texte raccourci, c'est-à-dire substituer au texte d'origine un texte condensé plus économique à manipuler, mais qui conserve les avantages informatifs du texte primitif. D'autre part, on souhaite disposer d'une espèce de sommaire du texte de départ, d'une représentation qui doit orienter sur le contenu de celui-ci sans prétendre se substituer à lui. Ainsi,

(5.3) Activité volcanique et situation des habitants dans l'île de Saint-Vincent.

peut être considéré comme une condensation de type "sommaire" de (5).

Considérons à nouveau le texte (5) pour envisager autrement la contraction. Supposons qu'à la suite de fâcheuses erreurs de transmission dans la graphie du texte, il soit reproduit sous la forme suivante :

(5.4) U calm appare règn au volca d la Soufrière d l'îl d Saint-Vincent. L'évacuat d villges situé à prximit dns l nor de l'îl s poursui cpendnt. L experts rest en effe inquie. Ils n save ps si c calm marq l fin d l'alert ou annonc un érupt prochai. L qlque quinz mil person q ont djà été évacué dns l sud d l'îl n regagnero ps lrs villge avnt au moins u moi.

Sans être particulièrement doué en décryptage, il est assez facile de suppléer aux manques. Or (5.4), qui présente 348 caractères ou espaces, soit 87 de moins que (5) qui en compte 435, peut être, lui aussi, considéré comme condensé par rapport à (5).

Chacune des condensations (5.1) à (5.4) illustre une possibilité différente de condensation de (5). Or entre (5.1) à (5.3) d'une part et (5.4) d'autre part, il existe un clivage important : la condensation de type (5.4) peut être obtenue sans tenir compte de conditions pragmatiques portant sur l'utilisation de l'information véhiculée par (5); par ailleurs, il est en principe possible d'obtenir (5) à partir de (5.4), en appliquant des règles déductibles du système linguistique sous-jacent, ce qui n'est pas le cas pour les autres condensations. Nous dirons que (5.4) est une condensation interne au système linguistique sous-jacent, alors que (5.1) à (5.3) sont ce que, faute de mieux, nous pouvons appeler des condensations "larges", condensations dans lesquelles des facteurs externes au système linguistique jouent un rôle essentiel. Condensation large et condensation interne peuvent se combiner :

(5.5) L calm q règn au volc d la Soufrière (îl d Saint-Vincent) inquièt l experts, q s demand s'i n'annonc ps un érupt prochai; l'évacuat d habitants s poursui dnc.

A un certain niveau d'abstraction, toutes les condensations obéissent à un même schéma : un texte de départ est associé à un texte d'arrivée par un certain processus, le texte d'arrivée étant plus court, par rapport à un ou plusieurs facteurs, que le texte de départ; le texte de départ est ainsi amputé sur un ou plusieurs plans différents. Cette amputation ne s'effectue pas au hasard : on élimine ce qui est en trop mais restituable, ou en trop mais secondaire. C'est ce que le symbole du C.I.L.N. veut représenter : les informations du texte d'entrée arrivent en parallèle dans une machine à condenser et, après traitement, donnent lieu à une sortie unique.

Il est vite apparu que la condensation large et la condensation interne ne peuvent - au moins pour l'instant - ni être étudiées avec les mêmes outils, ni prétendre à la même exactitude. Si le système linguistique sous-jacent, ou, tout au moins, certains aspects d'un système linguistique sous-jacent, peuvent être notés et manipulés avec un degré élevé d'explicitation,

les facteurs qui interviennent dans la condensation large - connaissance des conditions d'énonciation, du contexte, des visions du monde, etc. - restent largement rebelles à une formalisation; dans ce cas, la boîte de la machine à condenser, celle qui associe les entrées en parallèle à la sortie unique dans le symbole du C.I.L.N., a toutes les chances de rester noire longtemps encore. En revanche, pour certains cas de la condensation interne, il est déjà possible d'arriver à un degré d'explicitation tel qu'un ordinateur peut exécuter le travail de la machine à condenser.

Le fait que la problèmatique générale de la condensation soit susceptible d'être posée à différents niveaux de formalisation et qu'elle tienne compte soit exclusivement de facteurs internes au système linguistique, soit du système linguistique et d'autres facteurs, a conduit à traiter cette problématique en trois programmes séparés, que nous abordons aujourd'hui en parallèle, bien qu'il y ait des rapports entre eux.

Le programme Contraction de texte se propose d'étudier une machine à condenser humaine, qui reçoit en entrée un texte en langue naturelle et produit en sortie un nouveau texte en langue naturelle (cf dans ce cahier : "Problématique de la contraction de texte", p. 13). La machine est humaine car fortement tributaire d'une large zone d'intuition chez l'être humain qui effectue le travail de condensation. Par ailleurs, les textes d'entrée et de sortie sont destinés à être traités par un être humain.

Le programme Linguistique et langages documentaires se propose d'étudier les langages documentaires en fonction des caractéristiques des langues naturelles. Son axe central est l'élaboration du système documentaire Vercingétorix; celui-ci doit intégrer des langages documentaires de complexité syntaxique croissante (cf ici même "Le système documentaire Vercingétorix I" p. 57). Ce programme, comme le précédent, relève de la condensation large, mais, à la différence du précédent, son objectif est "une condensation de type sommaire" et non une condensation de substitution. La machine à condenser de Vercingétorix reste humaine. En outre, son entrée est toujours un texte en langue naturelle, ou au moins la formulation linguistique qui permet d'appréhender un document autre qu'un texte, mais sa sortie est un texte susceptible d'être compris aussi bien par l'être humain que par la machine. Cette machine à condenser reste donc tributaire de l'intuition de l'être humain qui reçoit les documents et les indexe dans le cadre du système documentaire Vercingétorix; mais les contraintes sur la sortie d'une part et l'environnement d'utilisation du langage documentaire d'autre part - en particulier l'assistance

par l'ordinateur à l'indexation - exigent de ce programme un niveau de formalisation plus élevé que le précédent.

Le programme sur la Codification économique du langage, enfin, est le plus exigeant au niveau de la formalisation. Il s' ensuit que son domaine d'application est plus restreint, en ce sens qu'il vise une condensation interne au système linguistique sous-jacent, et encore, dans la situation présente, limitée à un aspect très partiel de ce système : la redondance introduite par l'organisation du système phonologique (cf "Un théorème sur l'équivalence de la valeur de H entre langages", p. 97). La machine à condenser de ce programme doit être entièrement explicitée, elle reçoit en entier un texte en langue naturelle et produit en sortie, sans intervention humaine, une suite de symboles binaires, la plus réduite possible, à partir desquels il doit être possible de restituer le texte en langue naturelle de l'entrée.

Voilà donc le domaine et les axes de recherche,tels qu'ils sont aujourd'hui perçus,sur lesquels le C.I.L.N. se propose de continuer à travailler. Condenser recevra au fur et à mesure de leur élaboration - c'est-à-dire souvent sans un ordonnancement rhétorique impeccable-les fruits des travaux du C.I.L.N.; tâtonnements et rectifications sont donc à prévoir. Que les termes de cette introduction soient donc compris comme l'expression d'un véritable projet, et non comme le chapitre introductif que l'on a immanquablement rédigé alors que les conclusions sont déjà tirées. Travail en élaboration, toutes les issues restent ouvertes, y compris la possibilité de le formuler autrement et/ou de constater l'impossibilité de le traiter.

C.I.L.N.